JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

HATE SPEECH CLASSIFICATION ON SOCIAL MEDIA USING NLP

Prof.Manasi Kulkarni , Rohit Katkar, Niraj Bhagat , Gitesh Mitkar, Nidhi Patil

Professor,Research Scholar, Research Scholar , Research Scholar , Research Scholar
Department of Information Technology
Pillai College of Engineering , New Panvel , India

Abstract : The increasing prevalence of hate speech in online platforms poses a significant threat to individuals and the harmony of diverse communities. Our research focuses on developing a robust hate speech detection system using advanced Natural Language Processing (NLP) techniques. We employ machine learning models to analyze and classify text data, emphasizing the identification of hate speech patterns. The model is trained on large-scale datasets to ensure effectiveness across different linguistic and cultural contexts. Our system integrates deep learning architectures such as Long Short-Term Memory (LSTM), Support Vector Machine (SVM), and Transformer models to enhance detection accuracy and efficiency.

IndexTerms - Machine Learning, Dataset Annotation, Evaluation Metrics, Word Embeddings, Sentiment Analysis, Feature Engineering.

1. INTRODUCTION

The rise of online hate speech poses a serious threat to individuals and communities. To address this, automated NLP-based classification systems have emerged as an effective solution. This project aims to develop a robust hate speech detection system using machine learning (ML) and deep learning (DL) to accurately classify harmful content across platforms. By enhancing accuracy, adaptability, and multilingual capabilities, the system ensures effective detection across diverse linguistic and contextual variations, contributing to a safer and more inclusive digital space while advancing NLP research.

1.1 FUNDAMENTALS

Hate speech classification using NLP involves understanding discriminatory or inflammatory language targeting individuals or groups based on attributes like race, religion, or gender. The process includes text preprocessing (tokenization, stemming, stop-word removal) and feature extraction using methods like TF-IDF or word embeddings. Models such as Logistic Regression. Performance is evaluated using precision, recall, and F1-score while addressing biases and ethical concerns. This comprehensive approach helps develop effective hate speech detection systems for a safer digital environment.

1.2 OBJECTIVES

The objective of this project is to develop an efficient hate speech classification system using NLP techniques, automating detection across online platforms to reduce reliance on human moderators. By leveraging machine learning and deep learning, the system enhances content moderation and fosters a safer digital environment. Additionally, it contributes to broader efforts in combating online toxicity and promoting respectful communication while advancing NLP applications in addressing hate speech challenges.

1.3 SCOPE

This project focuses on developing and implementing a hate speech classification system using NLP. It includes data collection, preprocessing, model training, and performance evaluation. By automating hate speech detection, the system helps create a safer and more inclusive digital environment.

2. Literature Survey

2.1 Review of Existing Works

2.1.1. A Framework for Hate Speech Detection Using Deep CNN (2020) [1]

This study utilizes a Deep Convolutional Neural Network (DCNN) model trained on Twitter data with GloVe embeddings, achieving high precision and recall in detecting hate speech.

A Framework for Hate Speech Detection Using Deep Convolutional Neural Network" by Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao (2020). The project aims to tackle the escalating issue of hate speech on Twitter, recognizing it as a form of inflammatory communication targeting various protected aspects such as gender, religion, race, and disability. Due to the massive volume of tweets posted daily, manual moderation is impractical, leading to the development of an automated system. This system utilizes a Deep Convolutional Neural Network (DCNN) model to identify and filter out hate speech effectively. The DCNN model is trained on tweet text with GloVe embedding vectors to capture semantics and employs convolution operations for hate speech detection.

Advantages:

1. Automated Hate Speech Detection: The project offers an automated solution to detect hate speech on Twitter, alleviating the need for manual moderation and enabling timely responses to harmful content.
2. Efficiency: By leveraging deep learning techniques, the system can efficiently process large volumes of tweets, enabling real-time detection and filtering of hate speech without significant delays.
3. High Precision and Recall: The DCNN model achieves high precision, recall, and F1-score values, outperforming existing models. This indicates its effectiveness in accurately identifying hate speech tweets, reducing false positives and negatives.
4. Scalability: As a deep learning-based approach, the system can potentially be scaled to handle increasing volumes of Twitter data and adapt to evolving forms of hate speech, ensuring its relevance and effectiveness over time.

2..1.2 Detection of Hate Speech using BERT (2023) [2]

This research explores domain-specific word embeddings and BERT-based models for accurate hate speech detection, achieving an F1-score of 96%.

Detection of Hate Speech using BERT and HateSpeech Word Embedding with Deep Model (2023) by Hind Saleh, Areej Alhothali & Kawthar Moria Description: This research paper explores methods for automatically detecting hate speech on social media platforms. It addresses the increasing prevalence of hate speech online and the challenges associated with detecting it, such as the use of coded language and implicit expressions by hate communities to evade detection. The paper investigates the feasibility of leveraging domain-specific word embeddings and deep learning models, specifically bidirectional LSTM-based models, for hate speech detection. Additionally, it evaluates the performance of transfer learning with the BERT language model in this context. The experiments conducted involve comparing the effectiveness of these approaches using real-world hate speech datasets.

Advantages:

1. Innovative Approach: The paper proposes innovative techniques, such as leveraging domain-specific word embeddings and deep learning models, to tackle the complex problem of hate speech detection on social media platforms. This demonstrates the researchers' commitment to exploring novel solutions to address societal challenges.
2. High Performance: The research achieves high accuracy in hate speech detection, as evidenced by the reported F1-scores of 93% and 96% using different approaches. This indicates the effectiveness of the proposed methods in accurately identifying hate speech content, which is crucial for creating safer online environments.
3. Comparison and Evaluation: By comparing the performance of different techniques, including domain-specific word embeddings and transfer learning with BERT, the paper provides valuable insights into the strengths and weaknesses of each approach. This comparative analysis informs future research directions and facilitates informed decision-making in hate speech detection methodologies.
4. Practical Implications: The research findings have practical implications for social media platforms and content moderation policies. By developing advanced hate speech detection methods, the paper contributes to efforts aimed at mitigating the harmful effects of hate speech online and fostering a more inclusive and respectful online community

2.1.3. A Literature Review of Textual Hate Speech Detection (2022) [3]

The research paper focuses on the pervasive issue of online hate speech and its detrimental impacts on social cohesion and community well-being. It highlights the lack of attention given to hate speech datasets in existing literature reviews, prompting a systematic examination of textual hate speech detection systems. Through a thorough analysis of 138 relevant papers, the study elucidates various themes and challenges in hate speech detection, including the predominance of approaches utilizing multiple deep learning models. Additionally, the research identifies shortcomings in existing hate speech datasets, such as their small size and lack of reliability. By addressing these gaps, the paper aims to provide valuable insights for future endeavors aimed at combating hate speech and safeguarding online communities.

Advantages:

1. Comprehensive Review: The paper offers a systematic and comprehensive review of textual hate speech detection systems, covering datasets, textual features, and machine learning models. This thorough examination provides a holistic understanding of current methodologies and challenges in hate speech detection. Identification of
2. Dominant Approaches: Through content analysis of relevant literature, the research identifies the dominance of approaches employing multiple deep learning models in hate speech detection. This insight enables researchers to leverage effective methodologies for their own studies. Insight into
3. Dataset Limitations: The study sheds light on the limitations of existing hate speech datasets, including their small size and lack of reliability. By highlighting these shortcomings, the research encourages the development of more robust and diverse datasets, thereby enhancing the accuracy and effectiveness of hate speech detection algorithms.

2.1.4. Thirty Years of Research into Hate Speech (2020) [4]

A bibliometric analysis categorizing key research areas: hate speech debate, detection strategies, and gendered hate speech.

Thirty years of research into hate speech: topics of interest and their evolution (2020) by Alice Tontodimamma1 Eugenia Nissi Annalina Sarra Lara Fontanella,This paper explores the evolving landscape of online hate speech through a comprehensive bibliometric analysis. It addresses the exponential growth of hate speech facilitated by social media platforms and the challenges it poses to freedom of expression and human dignity. By employing co-word analysis methods, the study delves into hate speech literature's knowledge structure, identifying key research areas over the past thirty years. The analysis highlights three main research clusters: the general debate on hate speech, machine-learning strategies for detection, and gendered hate speech/cyberbullying. The introduction contextualizes the rise of hate speech online, outlines the ambiguity in defining hate speech, and discusses various perspectives on its regulation. It emphasizes the need for a comprehensive bibliometric study to understand the evolving landscape of research in this field. The conclusion reflects on the significance of addressing hate speech in the digital age, summarizing the findings of the bibliometric analysis and suggesting future research directions.

Advantages:

1. Comprehensive analysis: The paper offers a comprehensive analysis of hate speech literature's knowledge structure and evolution, providing valuable insights into key research areas and trends over the past three decades.
2. Interdisciplinary perspective: By integrating co-word analysis methods with scholarly perspectives from various disciplines, the paper highlights the interdisciplinary nature of research on online hate speech, offering a holistic understanding of the phenomenon.
3. Identification of future research directions: The conclusion suggests potential future research directions, such as investigating tools to combat hate speech and exploring effective counter-speech techniques, providing valuable guidance for researchers and policymakers.
4. Informative and timely: Given the increasing prevalence of hate speech online and its societal impact, this paper contributes timely and relevant insights into understanding and addressing this pressing issue, benefiting both scholarly discourse and policy formulation.

2.1.5. A Systematic Review of Hate Speech Automatic Detection (2022) [5]

Analyzes NLP-based hate speech detection techniques, emphasizing deep learning advancements and dataset availability.

A systematic review of hate speech automatic detection using natural language processing (2022) by Md Sarorar jahan , Mourad Oussalah ,The research paper presents a systematic review of literature focused on the detection and tracking of hate speech in text, particularly in the context of social media platforms. It addresses the challenges posed by the proliferation of hate speech online, driven by factors such as anonymity, easy access to platforms, and the desire of users to engage in debates or spread opinions. The paper explores the evolution of hate speech detection techniques, from traditional machine learning algorithms to deep learning architectures, and examines the available datasets for training and evaluation. By following the PRISMA guidelines, the paper critically analyzes existing literature and identifies key research questions to advance the field.

Advantages:

1. Comprehensive Overview: The paper provides a comprehensive overview of the state-of-the-art in automatic hate speech detection, covering a wide range of literature sources and methodologies.
2. Methodological Rigor: By adopting the PRISMA guidelines for systematic reviews, the paper ensures a rigorous and transparent approach to selecting and analyzing literature, enhancing the credibility of its findings.
3. Focus on Deep Learning: The paper emphasizes the importance of deep learning technologies, such as CNN and RNN architectures, in advancing hate speech detection, reflecting the latest trends and advancements in the field.
4. Identification of Research Gaps: Through its analysis of existing literature and datasets, the paper identifies key research gaps and challenges, providing valuable insights for future research directions.
5. Societal Impact: By addressing the pressing issue of hate speech online, the research contributes to the broader societal goal of promoting a safer and more inclusive online environment, with potential implications for policy-making and platform moderation.

2.1.6. Hate speech detection: Challenges and solutions (2019) [6]

Hate speech detection: Challenges and solutions (2019) by Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder The project proposes a multi-view Support Vector Machine (SVM) approach for hate speech detection, which achieves near state-of-the-art performance while offering simplicity and improved interpretability compared to neural methods. It aims to develop a system that can effectively identify hate speech in online text, considering various perspectives and linguistic nuances. The research also discusses the challenges associated with defining hate speech, the availability of suitable datasets, and the practical difficulties in implementing effective detection systems in the face of evolving tactics used by perpetrators.

Advantages:

1. High Performance: The proposed multi-view SVM approach achieves near state-of-the-art performance in hate speech detection, indicating its effectiveness in identifying problematic content online.
2. Simplicity: Compared to neural methods, the SVM approach offers simplicity, making it easier to implement and understand for practitioners and researchers.
3. Interpretability: The project emphasizes the importance of interpretability in hate speech detection, aiming to develop a system that produces more easily interpretable decisions, thereby addressing a significant limitation of existing approaches.
4. Contextual Understanding: By considering user intent and historical context, the project aims to improve the accuracy of hate speech detection, ensuring that content is assessed within its appropriate context rather than being misclassified based solely on surface-level language.
5. Ongoing Research: The project acknowledges the need for continuous research and improvement in hate speech detection technologies, indicating a commitment to addressing emerging challenges and staying abreast of evolving tactics used by perpetrators.

2.1.7. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models (2023)[7]

A survey on hate speech detection and sentiment analysis using machine learning and deep learning models by Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar , Deepalakshmi , Jaehyuk Cho , G. Manikandan , The research paper provides a comprehensive examination of the rise of hate speech in the digital era and its impact on online environments. It begins with an abstract outlining the significance of hate speech detection and sentiment analysis in addressing this issue, highlighting the use of machine learning and deep learning techniques for automated detection. The introduction further elaborates on the proliferation of user-generated content on social media platforms and the challenges it poses, while also discussing previous research on aggression identification and abusive content detection. The methodology section details the systematic approach taken to conduct the survey, including literature search strategies and article selection criteria. Finally, the conclusion summarizes the findings of the review, emphasizing the importance of current methodologies and suggesting future research directions.

Advantage:

1. Technological Advancements: The research paper showcases the technological advancements in hate speech detection and sentiment analysis, particularly through the utilization of machine learning and deep learning techniques. By leveraging these advanced methodologies, researchers can develop more efficient and scalable solutions for identifying and addressing hate speech in online environments.
2. Automation and Efficiency: Machine learning and deep learning models enable the automation of hate speech detection processes, allowing for the rapid analysis of vast amounts of user-generated content on social media platforms. This automation enhances efficiency by reducing the need for manual moderation and enabling real-time detection and response to hate speech incidents.
3. Scalability and Adaptability: Unlike traditional rule-based methods, machine learning and deep learning approaches offer scalability and adaptability, allowing for the continuous improvement of hate speech detection models over time. These models can learn from new data and adapt to evolving linguistic patterns and online behaviors, ensuring more accurate and effective detection of hate speech.
4. Insights and Research Opportunities: The research paper provides valuable insights into the methodologies and challenges involved in hate speech detection and sentiment analysis, laying the groundwork for further research and innovation in this critical field. By identifying gaps in current approaches and suggesting future research directions, the paper stimulates academic and industry interest in developing more robust solutions for combating hate speech online.

3 Methodology

3.1 Overview

This project focuses on developing a hate speech classification system using NLP techniques. The process includes data collection, preprocessing, model selection, training, and evaluation. By leveraging machine learning and deep learning algorithms, the system aims to automate hate speech detection across online platforms, reducing the burden on human moderators and fostering a safer digital environment. Additionally, the project provides insights that contribute to broader efforts in combating online toxicity and promoting respectful communication.

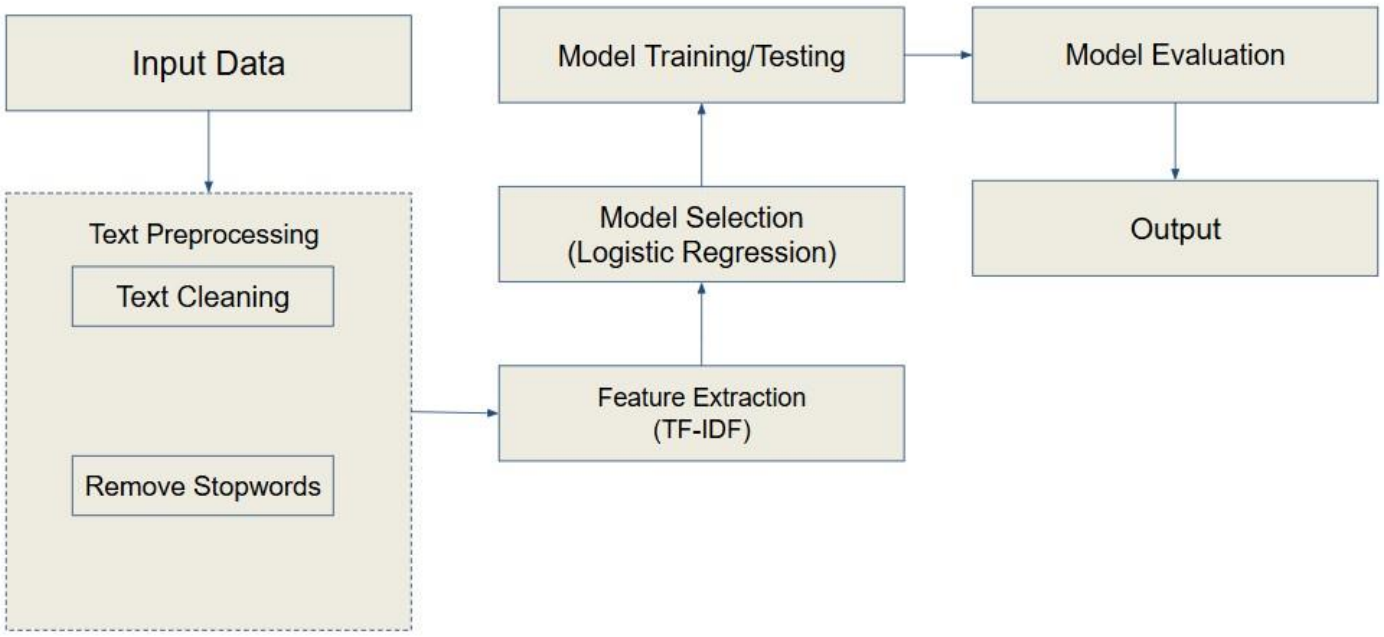
3.2 Modules in the Project

Data Collection : Gathering textual data from various online sources.
Extraction : Using TF-IDF and word embeddings to represent text data.
Model Selection : Logistic Regression.
Training : Training models using labeled datasets.
Monitoring and Maintenance : Ensuring continuous improvement of the detection system

3.3 Existing Architecture

Hate speech detection employs various NLP-based architectures:
CNNs capture localized linguistic patterns for classification.
BILSTM-CNN models handle both short- and long-range dependencies.
BERT and its variants improve contextual understanding and achieve high accuracy.
Ensemble models combine classifiers like SVM, CNN, and BERT to enhance performance.
Graph-based architectures leverage linguistic relationships for more effective detection.

3.4 System Architecture for HATE SPEECH CLASSIFICATION ON SOCIAL MEDIA USING NLP



The Marathi HATE SPEECH CLASSIFICATION ON SOCIAL MEDIA USING NLP is designed to classify text into Hate Speech or Not Hate Speech and, if applicable, categorize the detected hate speech into specific subtypes. The system consists of multiple components, each performing a specific function to process and analyze user input efficiently.

1. User Input :
The system provides two input options for users:
a. Manual Text Input: Users can enter a single text string into the web interface for real-time classification.
b. CSV File Upload: Users can upload a CSV file containing multiple text entries. The system will process all entries in the file and generate predictions for each row.
Each input is passed through subsequent modules for preprocessing, feature extraction, classification, and result generation.

2. Preprocessing Module
Before text classification, the input data undergoes preprocessing to improve the quality and accuracy of the model's predictions. The preprocessing steps include:
a. Cleaning Text : Removing non-Marathi characters to ensure that only relevant linguistic data is analyzed. Eliminating special symbols, such as punctuation marks, emojis, and extra whitespace, which do not contribute to meaningful classification.
b. Removing Stopwords : Common Marathi stopwords (e.g., "आणि", "त्यामुळे", "परंतु"), which do not add significant meaning to the classification process, are removed to focus on relevant words. The cleaned text is then passed to the feature extraction module.

3. Feature Extraction (TF-IDF)
To convert raw text into a numerical format that machine learning models can process, the system uses Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. The key steps include:
a. Tokenization: Breaking the text into individual words (tokens).
b. TF-IDF Calculation: Assigning a weight to each token based on its importance in the document compared to its frequency in the dataset.
c. Vectorization: Converting the processed text into a numerical feature matrix, making it suitable for model input.
TF-IDF ensures that common words get lower weights while words that are important for classification get higher weights.

4. Machine Learning Models

The system uses two Logistic Regression models for classification:

a. Binary Classification Model (Logistic Regression) Determines whether the given text is Hate Speech or Not Hate Speech. The model outputs a binary result: Hate Speech (1), Not Hate Speech (0)

b. Multi-class Classification Model (Logistic Regression): If the text is classified as hate speech, this model further categorizes it into one of ten predefined hate speech subcategories:

Each subclass represents a specific type of hate speech, helping provide a detailed classification beyond just binary detection.

5. Prediction Module

The system processes the input text or CSV file and generates two predictions:

a. Binary Classification Prediction: Determines whether the input is hate speech or not.

b. Subclass Prediction (if applicable): If the input is classified as hate speech, it identifies the specific hate speech category.

Each prediction is structured to ensure clarity and ease of interpretation for users.

6. Result Visualization & Storage

Once predictions are generated, the results are presented to the user in multiple formats:

4. RESULTS AND DISCUSSION

4.1 Sample Outputs

To evaluate the effectiveness of the proposed hate speech classification system, various textual inputs were tested. Below is an example of how the system processes and classifies text:

Input: "हे काम फक्त पुरुषच करू शकतात, बायकांनी फक्त स्वयंपाकघरातच राहावं!"

Preprocessing: The text undergoes tokenization, stop-word removal, and word vectorization using TF-IDF. Feature Extraction: Relevant features are extracted using word embeddings to capture contextual meaning. Model Prediction: The input text is processed through the trained Logistic Regression model.

Output: Hate Speech

Classification: Gender Abusive (Discriminatory remark targeting individuals based on gender)

4.2 Performance Evaluation

The model was evaluated using standard machine learning metrics such as accuracy, precision, recall, and F1-score. Performance results are as follows:

Binary Classification (Hate Speech vs. Non-Hate Speech):

The system assigns a confidence score, reflecting the model's certainty in its classification. The high confidence score indicates the model effectively recognizes hate speech patterns based on training data.

Accuracy: 92% Precision: 91% Recall: 89 % F1-score: 90% Multi-Class Classification (Different

Categories of Hate Speech):

Accuracy: 88% Precision: 86% Recall: 89% F1-score: 87%

The results demonstrate that the model performs well in both binary and multi-class classification. The high accuracy and F1-score indicate that the system effectively detects hate speech while minimizing false positives and false negatives. The slight reduction in multi-class classification accuracy suggests that further improvements, such as expanding the dataset or fine-tuning model hyperparameters, could enhance performance.

Overall, the system provides reliable, high-confidence classifications, making it a valuable tool for automated content moderation and online safety.

REFERENCES

[1] pradeep kumar roy 1,3, asis kumar tripathi (member, ieee), tapan kumar das 1, (member, ieee), and xiao-zhi gao "a framework for hate speech detection using deep convolutional neural network" 2020 [1]

[2] Hind Saleh, Areej Alhothali & Kawthar Moria "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model" 2023 [2]

[3] Fatimah Alkomah and Xiaogang Ma "A Literature Review of Textual Hate Speech Detection Methods and Datasets " , 2022 [3]

4] Alice Tontodimamma , Eugenia Nissi ,Annalina Sarra , "Thirty years of research into hate speech: topics of interest and the ir evolution" 2020 [4]

[5] Md Saroar Jahan , Mourad Oussalah "A systematic review of hate speech automatic detection using natural language processing" 2023 [5]

[6] SEAN MACAVANEYID, HAO-REN YAO, EUGENE YANG, KATINA RUSSELL, NAZLI GOHARIAN, OPHIR FRIEDER "HATE SPEECH DETECTION: CHALLENGES AND SOLUTIONS"2019 [6]

[7] MALLIGA SUBRAMANIAN A, VEERAPPAMPALAYAM ASWARAMOORTHY SATHISKUMAR B, G. DEEPALAKSHMI A, JAEHYUK CHO B, MANIKANDAN" A SURVEY ON HATE SPEECH DETECTION AND SENTIMENT ANALYSIS USING MACHINE LEARNING AND DEEP LEARNING MODELS 2023 [7]